Northeastern University, Boston, MA

College of Engineering

Department of Civil and Environmental Engineering

CIVE 7381: Transportation Demand Forecasting and Model Estimation

Problem Set 1

From: **Nathan David Obeng-Amoako** (NUID: 002607282)

To: Professor Haris Koutsopoulos

Submitted on: Friday, September 13

Fall 2024

Table of Contents

List of Figures

List of DataFrames

List of Equations

Problems

Northeastern

Department of Civil and Environmental Engineering
400 Snell Engineering
360 Huntington Ave.
Boston, MA 02115

CIVE 7381 Transportation Demand
Problem Set #1
Due: Wednesday, September 18, 2024

A planning organization has collected data on **all** families (entire population!) vacationing in a particular area during the summer. The entire population consists of 200 families with the associated information contained in the file PS1-data.xls, which is available on Canvas Problem Set 1. The file contains the following information for each family.

Duration of vacation	*DUR*	1 if vacation is longer than a week 0 otherwise
City	*CITY*	1 if vacation is at a city 0 otherwise
Camping	*CAMP*	1 if camping during vacation 0 otherwise
Income	*INC*	Average disposable income of household to which traveler belongs (in dollars)
Expenditure	*EXPEND*	Average amount spent by the family during the vacation (in dollars)
Education	*EDUC*	Education, measured by number of years after high school of the highest degree in household

1. Using the entire population perform the following tasks
    a. Calculate the means and standard deviations for *EXP, DUR, INC,* and *EDUC*.
    b. Calculate the correlation and covariance matrix for these four variables.
    c. Plot histograms for these variables.
    d. Plot the scatter plot matrix for these variables.
2. Using the *sampling* function in Excel generate 5 random samples from the population, each of size 40.
    a. For each sample find the sample mean of the various attributes and show them in a table.
    b. Plot the frequency histogram of the sample mean values for Income.
    c. Find the average and variance of the sample mean for income (across the five samples). Compare the average to the true population mean and the variance to the variance calculated by the formula for the standard error.
3. Generate two new samples: Sample 1 is of size 100 (sample 100 observations) and Sample 2 consists of 175 observations. Repeat parts 1a, 1b and 1c. Comment on how the sample averages and histograms compare to the corresponding quantities in the entire population. Which sample is closer? Comment on the impact of the sample size.

4. Using the data from one sample generated in (2) and Sample 2 generated in (3) develop the confidence interval for the mean of each sample at the 95% level. Compare the corresponding intervals and comment on the effect of the sample size.
5. Using the data from Sample 2 (question 3), test the null hypothesis H_0: μ= \$25,000, against the alternative hypothesis H_1: $\mu \neq$ \$25,000, where μ is the population mean (make any assumption regarding the variance that you think appropriate).
6. Based on the results of the analysis in questions 2-4 what is the sample size you would recommend (obviously larger samples are more accurate but also more expensive)?
7. Based on your understanding of the data and the problem, which of the variables above you would use for developing a model to predict expenditure per family?
8. What other variables, not included in the data set, do you consider important in building a better model to predict expenditures per family?

You can use Excel to generate the various random samples. First make sure that the Analysis Tool add-inn is activated (add it by File → Options → Add-ins). This will allow us to use a Random Number Generator. There are different ways to draw a random sample from a data set. See for example:

WikiHow. How to Create a Random Sample in Excel, http://www.wikihow.com/Create-a-Random-Sample-in-Excel

SurveyMonkey Blog. How to Create A Random Sample in Excel, https://www.surveymonkey.com/mp/random-sample-in-excel/

Solution to Problem 1

The provided dataset was read into a Pandas DataFrame in Python and the descriptive statistics of all the variables obtained (**DataFrame 1**).

DataFrame 1: Descriptive Statistics for the Variables in the Dataset

```python
desc_stats = df.describe()

# Replace sample std with population std
desc_stats.loc['std'] = df.std(ddof=0)

desc_stats
```

	OBSERVATION	DUR	CITY	CAMP	EDUC	EXPEND ($)	INC ($)
count	200.000000	200.000000	200.000000	200.0	200.000000	200.000000	200.000000
mean	100.500000	0.170000	0.315000	0.0	1.950000	548.252099	24858.082151
std	57.734305	0.375633	0.464516	0.0	1.751428	142.106262	6544.023285
min	1.000000	0.000000	0.000000	0.0	0.000000	191.242391	6530.636433
25%	50.750000	0.000000	0.000000	0.0	1.000000	454.910772	20366.882217
50%	100.500000	0.000000	0.000000	0.0	1.000000	542.939203	24017.101686
75%	150.250000	0.000000	1.000000	0.0	3.000000	646.735277	29534.373943
max	200.000000	1.000000	1.000000	0.0	8.000000	908.651977	47858.148441

Problem 1a

From **DataFrame 1**, the mean and standard deviation for the variables **EXP**, **DUR**, **INC**, and **EDUC** have been summarized in **DataFrame 2.**

DataFrame 2: Mean and Standard Deviation of Select Variables

```python
[5] four_attributes = ['EXPEND ($)', 'DUR', 'INC ($)', 'EDUC']
    desc_stats[four_attributes].iloc[1:3]
```

	EXPEND ($)	DUR	INC ($)	EDUC
mean	548.25	0.17	24858.08	1.95
std	142.11	0.38	6544.02	1.75

Problem 1b

The correlation matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** was evaluated in Python and has been displayed in **DataFrame 3**.

DataFrame 3: Correlation Matrix

```
pd.set_option('display.float_format', '{:.5f}'.format)

corr_matrix = dF[four_attributes].corr()
corr_matrix
```

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	1.00000	-0.10310	0.59626	0.12096
DUR	-0.10310	1.00000	0.05616	0.11932
INC ($)	0.59626	0.05616	1.00000	-0.16138
EDUC	0.12096	0.11932	-0.16138	1.00000

The covariance matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** was evaluated in Python and has been displayed in **DataFrame 4**.

DataFrame 4: Covariance Matrix

```
[43] pd.set_option('display.float_format', '{:.2f}'.format)

cov_matrix = dF[four_attributes].cov()
cov_matrix
```

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	20295.67	-5.53	557274.69	30.26
DUR	-5.53	0.14	138.75	0.08
INC ($)	557274.69	138.75	43039437.94	-1858.96
EDUC	30.26	0.08	-1858.96	3.08

Problem 1c

Histograms for the selected variables have been illustrated in **Figure 1**.

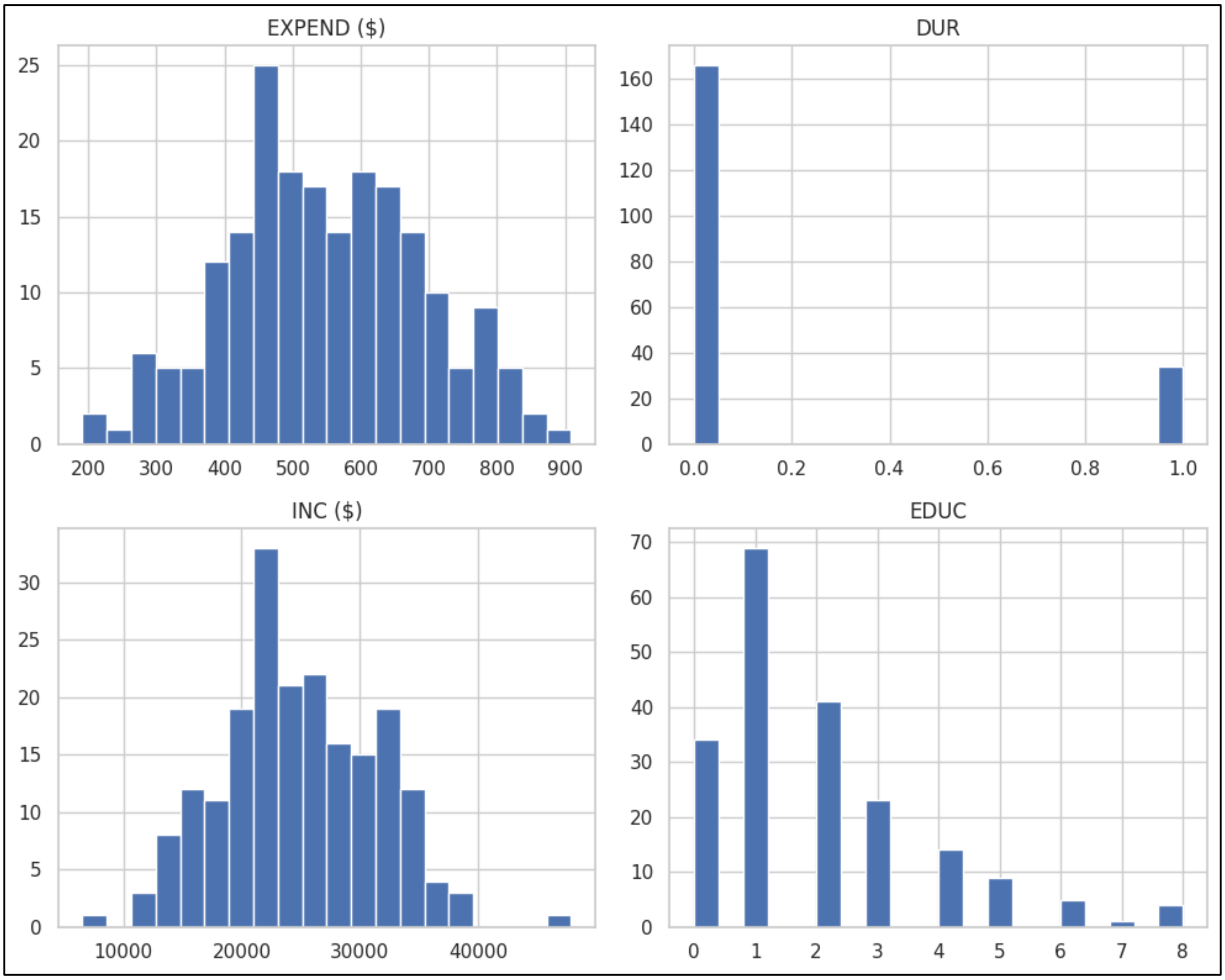

Figure 1: Histograms of EXPEND ($), DUR, INC ($), and EDUC distributions

Problem 1d

Scatter matrix for the selected variables has been illustrated in **Figure 2**. The leading diagonal shows a plot of the univariate histogram for each of the four attributes.

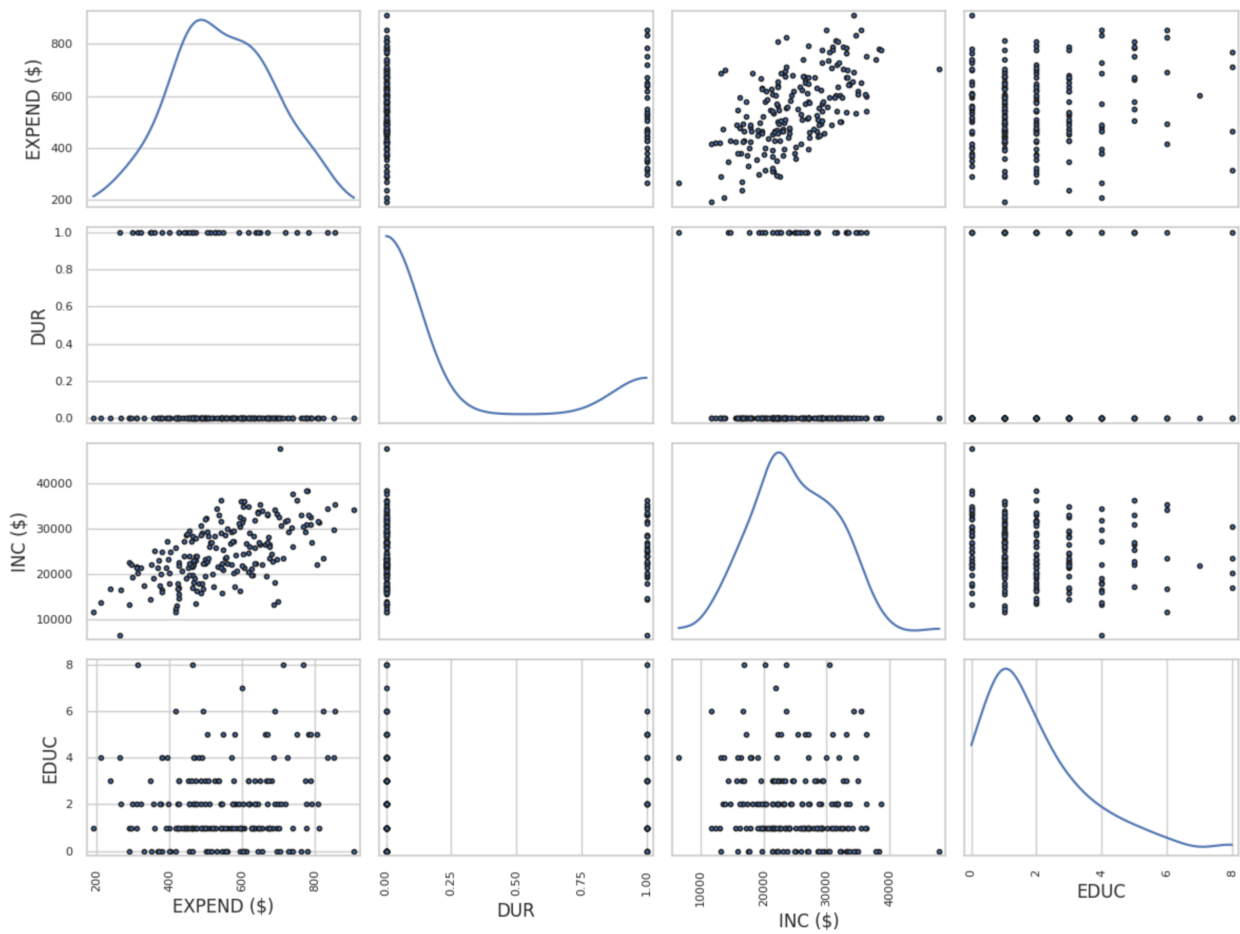

Figure 2: Scatter Matrix of EXPEND ($), DUR, INC ($), and EDUC

Solution to Problem 2

Problem 2a

The sample mean of each of the four selected variables for each of the five samples was calculated in Python and has been illustrated in **DataFrame 5**.

DataFrame 5: Sample Mean of Selected Attributes for 5 Random Samples

	EXPEND ($)	DUR	INC ($)	EDUC
Sample 1	534.70	0.23	25130.10	2.00
Sample 2	535.36	0.10	25395.58	1.98
Sample 3	570.87	0.15	24889.90	2.05
Sample 4	558.77	0.17	25281.92	1.98
Sample 5	565.64	0.17	24253.36	2.10

Problem 2b

The frequency histogram of the sample mean values for **INC ($)** is shown in **Figure 3**.

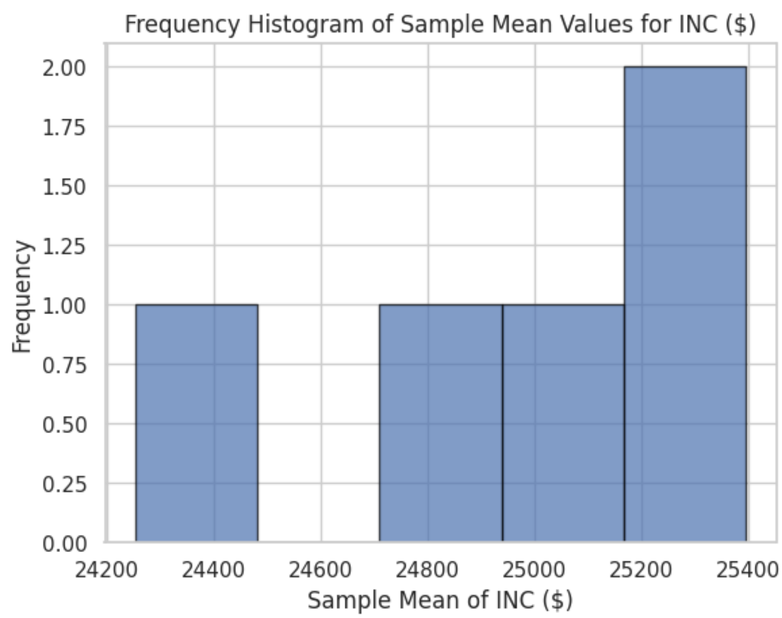

Figure 3: Frequency Histogram of the Sample Mean Values for INC ($)

Problem 2c

For the **INC ($)** attribute and across the five samples each of size 40,

- Average of sample means = **$ 24,990.17**
- Variance of sample means = **$ 606,334.90**

To recall,

- Population mean of **INC ($)** attribute = **$ 24,858.08**
- Population variance of **INC ($)** attribute = **$ 42,824,197.76**

From the standard error formula, the variance of the sample mean for the **INC ($)** attribute is given by the population variance divided by the number of samples, which is **$ 1,070,604.94**.

The mean of the sample means for the **INC ($)** attribute across the five samples closely matches the population mean. According to the Central Limit Theorem, the mean of the sampling distribution approaches the population mean as the sample size and the number of different samples increases.

There seems to be a significant discrepancy between the variance of the sample means for **INC ($)** and the theoretical variance as calculated using the standard error formula. This may be due to the small number of samples, which increases the chance of sampling variability and makes the empirical variance less reflective of the theoretical value. From the Central Limit Theorem, as the number of samples increases, the variance of sample means would converge more closely to the theoretical value predicted by the population variance.

Solution to Problem 3

The means and standard deviations of the four specified variables for the sample of size 100 was obtained from Python and the results displayed in **DataFrame 6**.

DataFrame 6: Mean and Standard Deviation of Select Variables (n = 100)

	EXPEND ($)	DUR	INC ($)	EDUC
mean	548.83	0.16	25120.86	2.10
std	145.83	0.37	6769.94	1.93

The means and standard deviations of the four specified variables for the sample of size 175 was obtained from Python and the results displayed in Error! Reference source not found..

DataFrame 7: Mean and Standard Deviation of Select Variables (n = 175)

	EXPEND ($)	DUR	INC ($)	EDUC
mean	548.19	0.17	25056.92	1.95
std	146.31	0.38	6575.54	1.72

The correlation matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** for the sample of size 100 was evaluated in Python and has been displayed in **DataFrame 8**.

DataFrame 8: Correlation Matrix (n = 100)

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	1.00000	-0.00593	0.60961	0.23744
DUR	-0.00593	1.00000	0.06594	0.10480
INC ($)	0.60961	0.06594	1.00000	-0.10608
EDUC	0.23744	0.10480	-0.10608	1.00000

The correlation matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** for the sample of size 175 was evaluated in Python and has been displayed in **DataFrame 9**.

DataFrame 9: Correlation Matrix (n = 175)

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	1.00000	-0.08061	0.62955	0.12071
DUR	-0.08061	1.00000	0.07355	0.17214
INC ($)	0.62955	0.07355	1.00000	-0.14554
EDUC	0.12071	0.17214	-0.14554	1.00000

The covariance matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** for the sample of size 100 was evaluated in Python and has been displayed in **DataFrame 10**.

DataFrame 10: Covariance Matrix (n = 100)

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	21482.39	-0.32	607942.39	67.37
DUR	-0.32	0.14	165.32	0.07
INC ($)	607942.39	165.32	46295056.33	-1397.23
EDUC	67.37	0.07	-1397.23	3.75

The covariance matrix for the variables **EXP**, **DUR**, **INC**, and **EDUC** for the sample of size 100 was evaluated in Python and has been displayed in **DataFrame 11**.

DataFrame 11: Covariance Matrix (n = 175)

	EXPEND ($)	DUR	INC ($)	EDUC
EXPEND ($)	21530.65	-4.47	609167.72	30.58
DUR	-4.47	0.14	183.33	0.11
INC ($)	609167.72	183.33	43486250.96	-1656.83
EDUC	30.58	0.11	-1656.83	2.98

Histograms for the selected variables for the sample of size 100 have been illustrated in **Figure 4**.

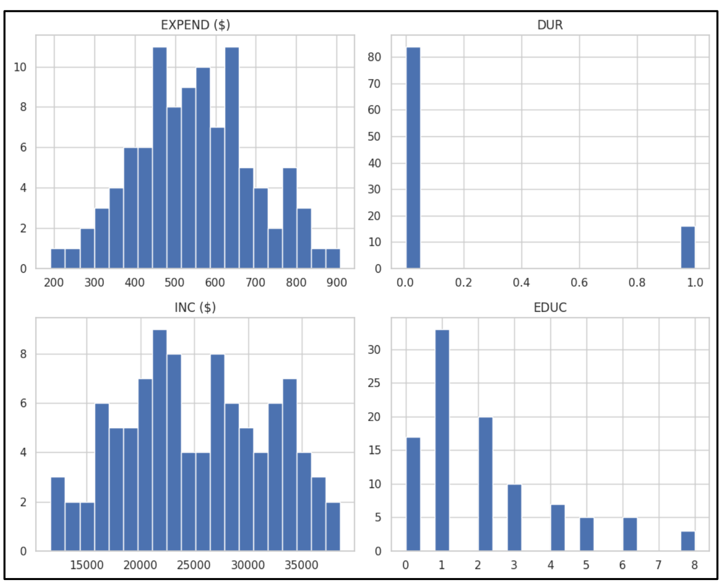

Figure 4: Histograms of EXPEND ($), DUR, INC ($), and EDUC distributions (n = 100)

Histograms for the selected variables for the sample of size 175 have been illustrated in **Figure 5**.

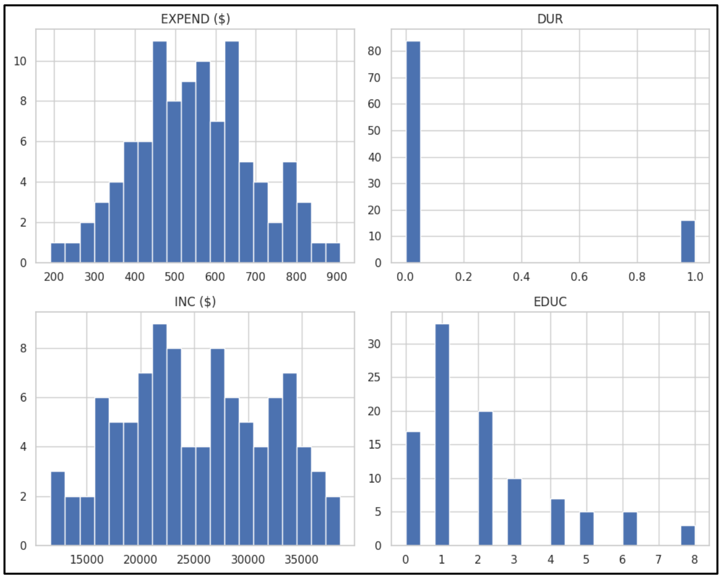

Figure 5: Histograms of EXPEND ($), DUR, INC ($), and EDUC distributions (n = 175)

Comments on the results:

In general, the bigger the sample size, the more closely the descriptive statistics match that of the population. Hence, Sample 2 gave more accurate results than Sample 1.

Solution to Problem 4

The confidence interval can be calculated using **Equation 1**.

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval
$\bar{x}$ = sample mean
z = confidence level value
s = sample standard deviation
n = sample size

Equation 1: Confidence Interval

The confidence interval at the 95% level (confidence level value = 1.96) for the four selected variables in the sample of size 40 is shown in **DataFrame 12**.

DataFrame 12: Confidence Interval (Sample Size 40)

	Mean	CI Lower Bound	CI Upper Bound
EXPEND ($)	564.70	524.90	604.50
DUR	0.23	0.09	0.36
INC ($)	25424.22	23533.56	27314.88
EDUC	2.12	1.54	2.71

The confidence interval at the 95% level (confidence level value = 1.96) for the four selected variables in the sample of size 175 is shown in **DataFrame 13**.

DataFrame 13: Confidence Interval (Sample Size 175)

	Mean	CI Lower Bound	CI Upper Bound
EXPEND ($)	545.33	524.01	566.66
DUR	0.17	0.12	0.23
INC ($)	24527.34	23574.36	25480.33
EDUC	1.97	1.71	2.24

With a sample size of 40, the 95% confidence intervals for the means tend to be wider, reflecting a higher degree of uncertainty (or higher margin of error) and less precision in estimating the population mean. In contrast, with a larger sample size of 175, the confidence intervals are narrower, indicating more precise estimates (or less variability) and greater confidence in the accuracy of the population mean. From **Equation 1**, larger sample size and smaller standard deviation, imply smaller margin of error, and hence, narrower confidence interval (higher precision).

Solution to Problem 5

The sample mean and standard deviation of the **INC ($)** attribute in the sample of size 175 were found to be $ 25,056.92 and $ 6,575.54, respectively. Given a hypothesized mean of $ 25,000.00 and using **Equation 2**, the t-statistic, **t**, is found to be **0.11451**.

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:

- $\bar{x}$ = the sample mean
- μ_0 = the hypothesized population mean
- s = the sample standard deviation
- n = the sample size

Equation 2: T-Statistic

Assuming a confidence interval of 95%, which corresponds to α = 5% (2.5% in each tail of the distribution), the critical t-value for a two-tailed test interpolated from the t-test table (**Appendix**) is around **1.97369**.

Because the absolute value of the t-statistic (0.11451) is less than the critical t-value (1.97369), I fail to reject the null hypothesis.

Assumptions made:

1. Population data is approximately normally distributed
2. The population variance found in Problem 1 was used.

Solution to Problem 6

There's always a trade-off between cost and accuracy. Smaller samples are more cost-effective but may be inaccurate. Larger samples are more accurate but can be expensive. As much as possible, I want the sample size to be parsimonious—just as large as necessary. That way, there's a balance in the trade-off. To help me decide, I developed a plot showing how sample size affects the accuracy of the mean and standard deviation. This has been illustrated in **Figure 6**.

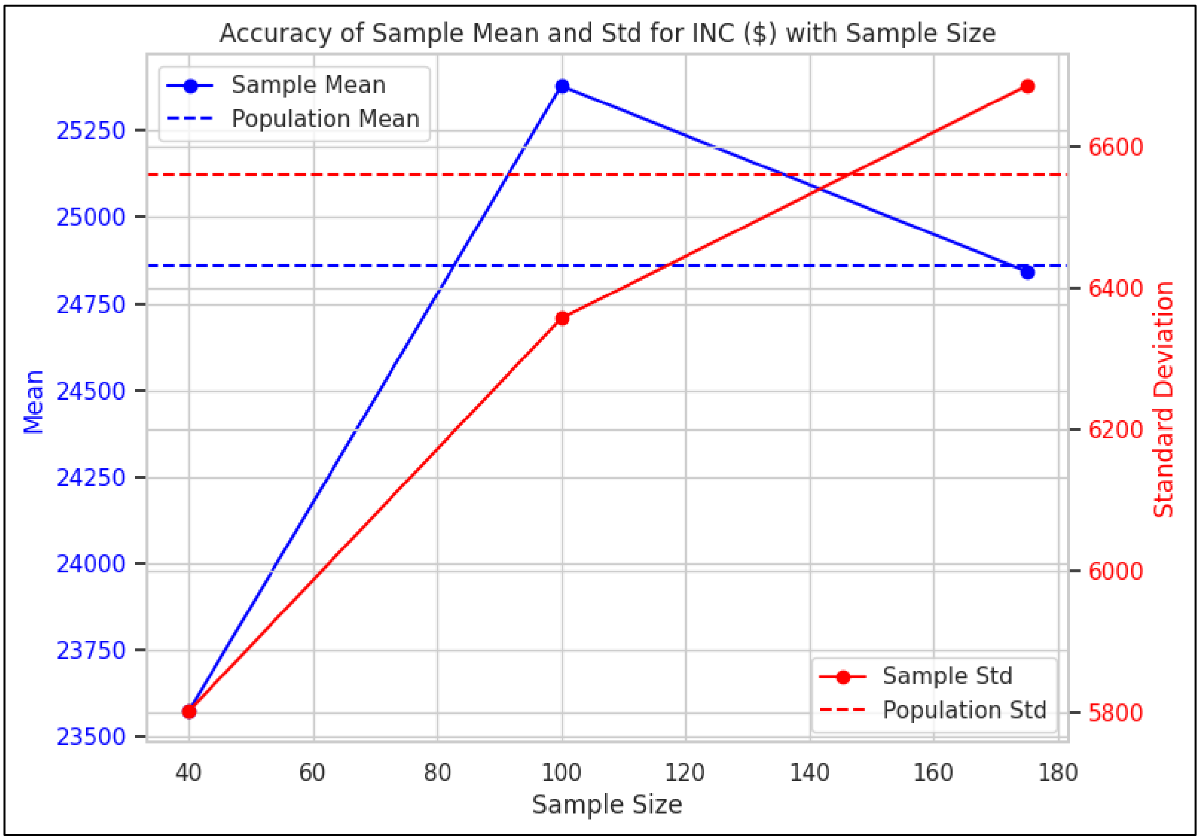

Figure 6: Accuracy of Sample Mean and Standard Deviation for INC ($) with Sample Size

From **Figure 6,** it is clear that the sample with size 175 provides the best results. However, the sample of size 100 provides results that are only slightly less accurate than the former. To balance cost and accuracy, I'll choose the **sample with size 100**.

Solution to Problem 7

All of the variables may be of help, to some degree, in predicting expenditure per family. However, judging based on the correlation matrix shown in **DataFrame 9**, the only other variable that exhibits a strong correlation with the expenditure variable is the income variable **INC ($)**. Hence, I'll use the income variable to predict expenditure per family.

Solution to Problem 8

Other variables that may prove important in predicting the expenditure of a family are:

1. family size
2. distance between home city and vacation city
3. nationality (this influences tax, visa requirement, etc.)

Appendix

The t-distribution to determine the critical value of the sample and compare it with the t-value to determine whether to reject the null hypothesis.

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										